

TITLE	Deliverable D2.2 Prototype
Document Type:	Report on prototype including user manual
WP/Task:	WP2
Document ID:	VIDI-02-20091231-D2.2
Version:	1.0
Date:	31.12.09
Status:	Final

Organisation:	Responsible partners: JSI
Authors:	Mitja Trampuš, Marko Grobelnik, Dunja Mladenić
Contributors:	Blaž Fortuna, Blaž Novak, Nenad Stojanović, Sinan Sen

Distribution:	PARTNERS
---------------	----------

Purpose of Document:	VIDI D2.2 Integrated and Evaluated VIDI system & System Manual
----------------------	--

Document History:	17.12.2009 outline of the deliverable, JSI 20.12.2009 First version of the deliverable, JSI 22.12.2009 Overall revision, JSI 28.12.2009 Adding section on Notifications, FZI 31.12. 2008 Final
-------------------	--

Integrated and Evaluated VIDI system & System Manual

VIDI Deliverable 2.2

31 December 2009

Table of contents

TABLE OF CONTENTS.....	3
EXECUTIVE SUMMARY	5
1. VIDI TOOLBAR FUNCTIONALITY.....	6
1.1. REAL-TIME NOTIFICATIONS.....	7
1.2. BROWSING SUGGESTIONS	7
1.3. TOPICAL ATLAS.....	7
1.4. TOPICAL TIMELINE.....	7
2. DEPLOYMENT MODES AND INSTALLATION	8
2.1. SERVER-SIDE DEPLOYMENT	8
2.2. CLIENT-SIDE DEPLOYMENT	8
3. USER MANUAL.....	10
3.1. GETTING STARTED.....	10
3.2. EXPRESSING AREAS OF INTEREST	11
3.3. OBTAINING BROWSING SUGGESTIONS.....	11
3.4. USING THE TOPICAL ATLAS	12
3.5. USING THE TOPICAL TIMELINE	14
3.6. DEFINING NOTIFICATION PATTERNS	15
4. LIVENETLIFE	17
5. TECHNICAL BACKGROUND	18
5.1. DATA ACQUISITION	18
5.2. DATA AGGREGATION AND AUGMENTATION	18
5.3. ANALYTIC MODULES	18
5.4. CLIENT SIDE AND CLIENT-SERVER COMMUNICATION	19
6. REFERENCES.....	20
APPENDIX A. SOURCE CODE ORGANIZATION	21

eParticipation Workprogramme

VIDI

Visualising the impact of the legislation by analysing public Discussions using statistical means

Project Reference No: EP-07-01-014

EXECUTIVE SUMMARY

This document describes the VIDI toolbar, i.e. the software component of the project. The toolbar was developed following the VIDI software architecture proposed in D2.1. It includes a data acquisition component (database access and Web crawling), a basic data handling component (data aggregation and data augmentation) and analytic modules (providing notifications, browsing suggestions, topical atlas and topical timeline).

The document targets three groups of readers:

- Software end users. The document includes a user's manual describing the use and functionality of the VIDI toolbar.
- Forum owners. One of the two possible modes of deployment for the VIDI toolbar is to install it on the server hosting the forum. The document provides instructions to this end.
- Developers. A technical overview of the architecture and algorithms is given in Section 5. In Appendix A, we describe the structure and organization of the source files comprising the VIDI toolbar. Further information is provided in the forms of comments within the source code.

Note that accompanying CD contains the source code of the presented software, as well as the installation instructions

1. VIDI TOOLBAR FUNCTIONALITY

As shown in Figure 1, the VIDI toolbar is used in parallel with the web forum of interest. It is composed of a *selection panel* in the upper part of the toolbar and three *action buttons* further down (as indicated by the two arrows in Figure 1). Using the selection panel, the user can select the part of the forum she might be interested in. Once the selection has been made, the action buttons provide access to the main VIDI toolbar functionalities, described in the following subsections. A more detailed description of the functionalities is available in VIDI Deliverable D2.1.

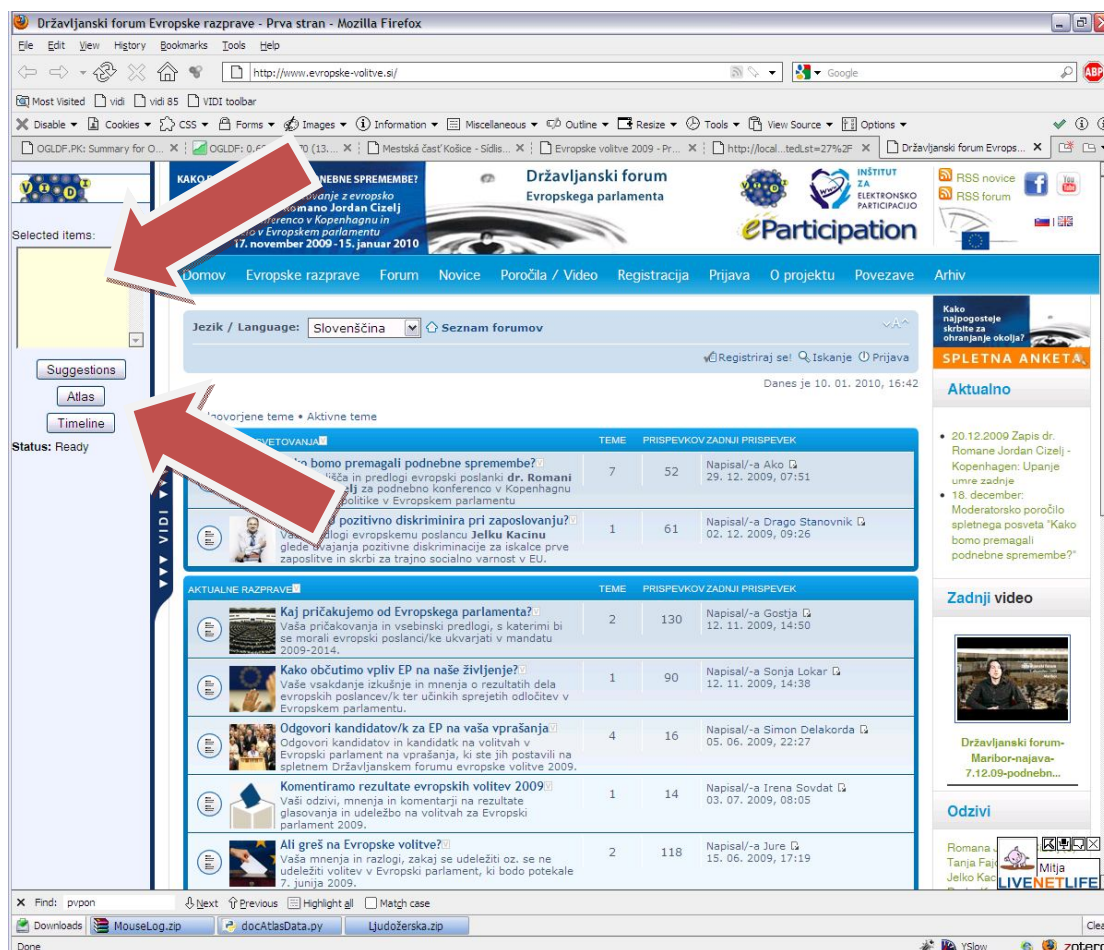


Figure 1: The toolbar deployed on INEPA's web forum. The selection panel and action buttons are marked with arrows.

1.1. Real-time Notifications

A discussion forum usually contains a lot of discussions and plenty of discussion topics. In some cases the user is only interested in a couple of discussion topics and wants to be alerted if certain discussion topics become more important or new facts are posted within new topics. In order to inform the user about these kinds of information the VIDI toolbar offers a user-driven real-time notification functionality. Once the user has specified his interest the notification system starts observing the discussion forum and as soon as the situation of interest happens to inform the user about the situation. Using this functionality the user is able to be informed about important changes within discussion forums without worrying about to miss important changes. See also section 3.6 for defining notification patterns.

1.2. Browsing Suggestions

The toolbar can, given a list of topics and threads of interest for the user, suggest further topics and threads on the current page that are similar to the selected ones and therefore potentially also interesting. For a more technical description, consult Section 3.1.2, "Link highlighting", in deliverable D2.1.

1.3. Topical Atlas

As a means of visually structuring a large subset of the forum, VIDI toolbar can plot all posts from a selection of topics and threads in a two-dimensional space. Each post is represented by a point and the points are arranged in such a way that the proximity of two points is roughly proportional to the similarity of the two corresponding posts. This way, clusters of points naturally form, representing subtopics of the part of the forum that is being analyzed. For each cluster, the user is able to determine the keywords (and with that, the topic).

See also section 3.1.3.1, "Semantic Space Visualization", in deliverable D2.1.

1.4. Topical Timeline

To determine the popularity of topics through time, the VIDI toolbar can automatically determine the relevant topics in a given subset of the forum and plot the number of posts on each of the topics as a function of time. The number of identified topics (and consequently their specificity) can be adjusted by the user.

See also section 3.1.3.2, "Canyon Flow Visualization", in deliverable D2.1.

2. DEPLOYMENT MODES AND INSTALLATION

Deliverable D2.1 envisioned the VIDI toolbar to be deployed as an Internet Explorer plugin. The upside of this approach is that it requires no involvement from forum owners. The downside is that the users have to run an installer in order to use the toolbar, which limits the dissemination potential of the toolbar. Another shortcoming of this approach is that it cannot support browsers other than Internet Explorer.

We have reconsidered the idea and consequently the toolbar has instead been developed using the GWT¹ platform. This means that the client side of the toolbar is written in javascript and can be deployed in two ways:

- A reference to the relevant javascript file can be inserted in the HTML template of the forum pages by the forum owner, making the toolbar available to all visitors without any involvement on their part.
- As an alternative, if the forum owner is not willing to include the toolbar, the user can still inject the relevant javascript into the page using bookmarklets².

Both methods are independent of the browser, as long as it has sufficiently strong support for javascript, which is the case with all major modern browsers. The javascript in question injects the toolbar's HTML code in the existing page, making the toolbar appear.

The rest of this section describes the "installation" process for each of the two methods.

2.1. *Server-Side Deployment*

The forum owner only needs to insert the following code in the <head> section of forum's HTML template:

```
<script type="text/javascript" id="gwt_vidi"
src="http://vidi.ijs.si/forumexplorerbar/forumexplorerbar.nocache.js"></script>
```

Nothing more is required. The toolbar will appear in hidden state (see Figure 3) for each user visiting the page.

2.2. *Client-Side Deployment*

If the user wishes to use the VIDI toolbar on a forum that does not include the above snippet of code in their pages, he should create a bookmark button pointing to the following URL:

¹ Google Web Toolkit: <http://code.google.com/webtoolkit/>

² <http://en.wikipedia.org/wiki/Bookmarklet>

```
javascript:(function(){var%20e=document.createElement('script');e.id='gwt_vidi';e.src='http://vidi.ijs.si/forumexplorerbar/forumexplorerbar.nocache.js?bookmarklet=1';document.body.appendChild(e);})();
```

For convenience, further instructions on how to create a bookmark button and a copy-paste-ready version of the above URL are available at <http://vidi.ijs.si/install.html>.

Once the bookmark button has been created, the user can use the VIDI toolbar by visiting a VIDI-supported forum of interest and pressing that button.

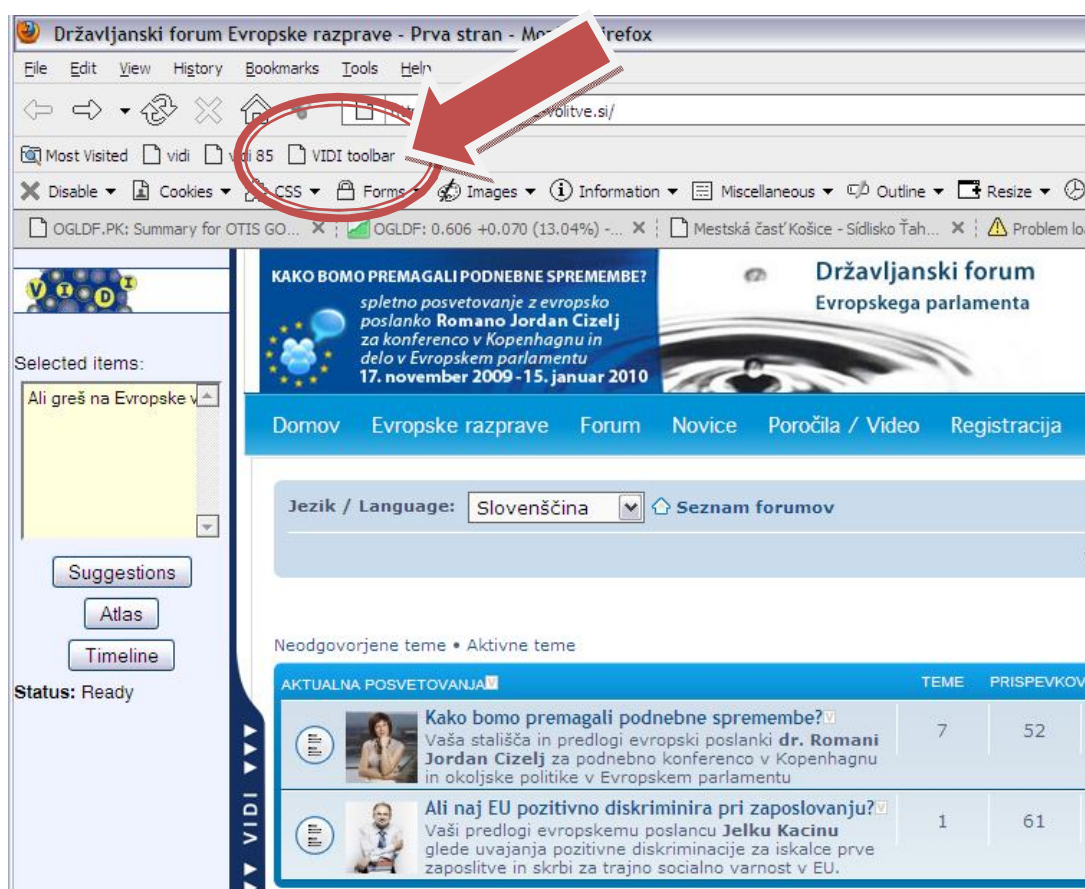


Figure 2: Client-side deployment with a bookmark button – example in Firefox.

3. USER MANUAL

3.1. Getting Started

To use the VIDI toolbar, navigate to a VIDI-supported forum. In the scope of this project, the following forums have been supported:

- Evropske volitve (Slovene)
<http://www.evropske-volitve.si/>
- MC Košice – Sídliisko Tahanovce (Slovak)
<http://mutah.tahanovce.sk:8080/mutah/web/sk/forums.jsp?id=50049>
- Political part of index.hu (Hungarian)
<http://forum.index.hu/Topic/showTopicList?t=9111313>

If the forum owner has installed the toolbar forum-wide, it will appear unobtrusively hidden in the left border of the page as shown in Figure 3. Clicking on the blue handle expands the toolbar, making it ready for use. If the toolbar does not appear automatically, not even in the hidden state, the forum owner has most likely not installed it. Please follow instructions in Section 2 to create a bookmark button in your own browser. Clicking the bookmark button will make the toolbar appear in expanded state.

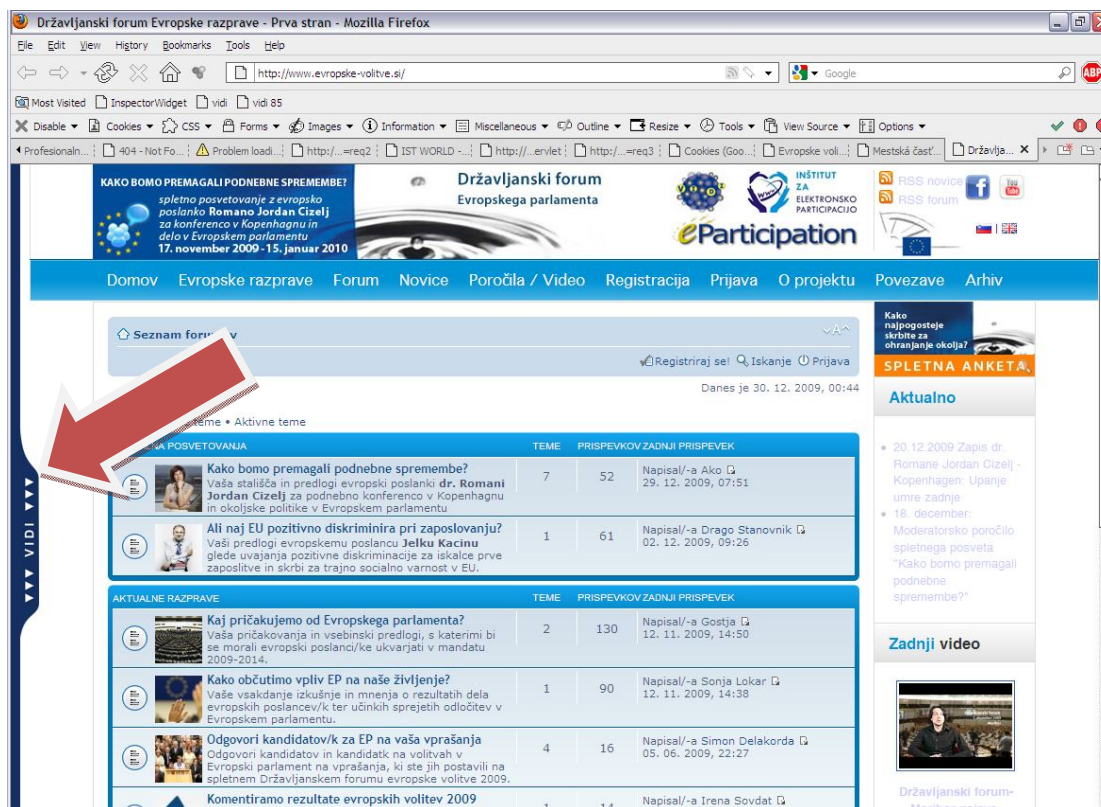


Figure 3: The toolbar in its hidden state.

3.2. Expressing Areas of Interest

In addition to the toolbar, there is another way in which the VIDI platform makes itself seen on the web page: Small icons appear next to each link that points to a discussion topic or a discussion thread, as illustrated in Figure 4. Click an icon to express interest in the corresponding forum section. The sections selected in this way are listed in the yellowish selection panel at the top of VIDI toolbar. Click an icon again to deselect the corresponding forum section.

Note: In deliverable D2.1, an automatic, implicit detection of user's browsing interests was foreseen as opposed to the explicit selection panel/icons combination. However, it has later been determined that users prefer tighter control over specifying their current interest as it may not have much to do with their past interests, especially for casual visitors to the site.



Figure 4: A close-up of the web page showing the small VIDI icons with which the user can express interest in the topic(s) of choice.

3.3. Obtaining Browsing Suggestions

To obtain suggestions on which forum threads might be of interest to you, first indicate your interest by selecting several threads as described in the previous section. Then, click the first button on the VIDI panel, "Suggestions". After a few seconds, links on the current page identified by the system as relevant to you will be marked with an orange icon; see Figure 5. The bigger the icon, the higher is the probability that the link is truly relevant.

To clear the suggestions, reload the page.

eParticipation Workprogramme VIDI

Visualising the impact of the legislation by analysing public Discussions using statistical means

Project Reference No: EP-07-01-014

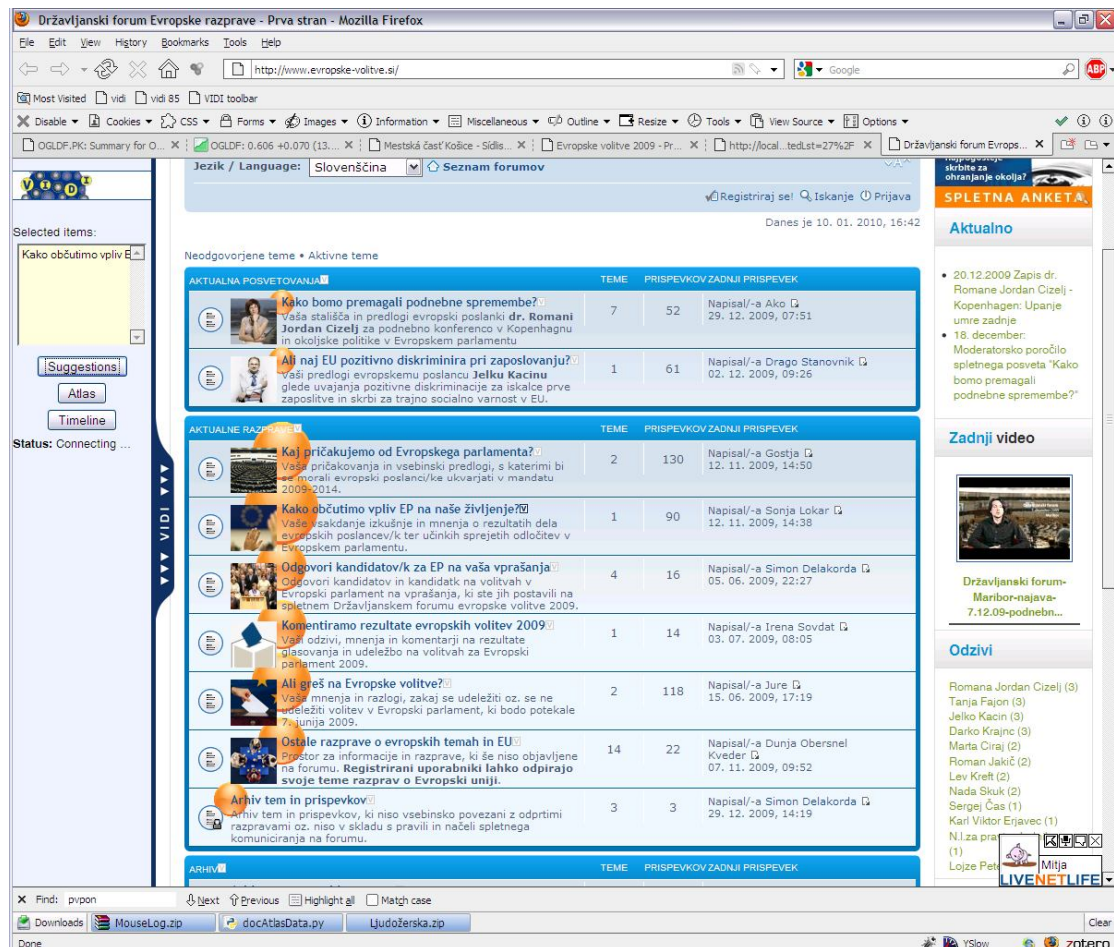


Figure 5: Browsing suggestions – the relevant links are marked by orange icons, their size proportional to link relevance.

3.4. Using the Topical Atlas

To see an "atlas" of all posts in a chosen subpart of the forum as illustrated in Figure 6, first indicate your interest by selecting several threads as described in section 3.2. Then, press the "Atlas" button on the VIDI toolbar.

The atlas appears in the middle of the window. Calculating all the data needed to display the atlas can take several minutes, so please be patient. To discard the atlas, click anywhere outside the popup area.

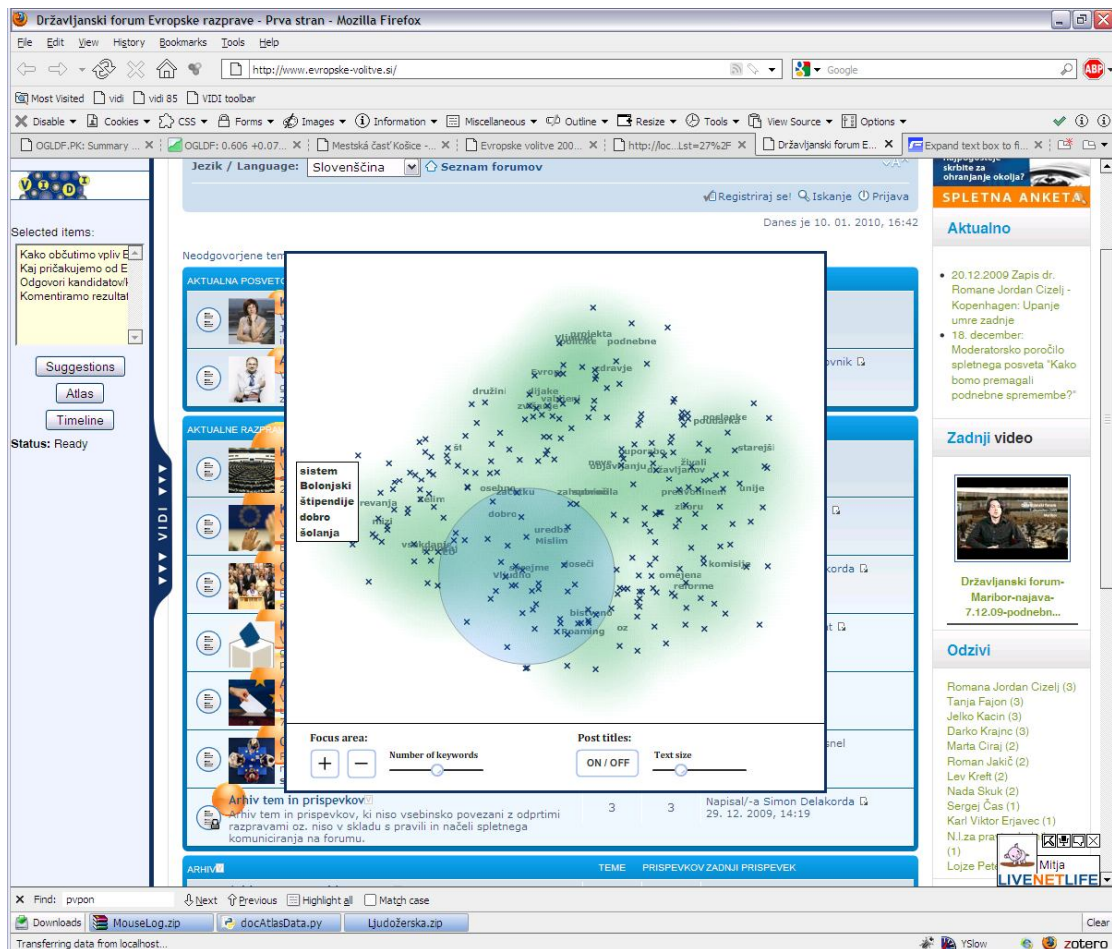


Figure 6: The "topical atlas" of a part of the forum. Similar posts are displayed close together, forming topical clusters.

The atlas chart comprises points (each representing a forum post) and keywords (each roughly describing the topic of its immediate neighborhood). To get further information about parts of the chart, move your mouse over it. The lightly blue shaded area under the mouse cursor is the *focus area* – posts within the focus area are summarized by a list of keywords that appears next to your mouse cursor.

To reduce the need to scan the whole chart with the focus area, some keywords are given in advance. Those appear in green in the background. The area of the chart of which they are representative is marked with a light shade of green.

Hovering the cursor over a forum post shows its subject as it appears on the forum (e.g. "Re: new anti-smoking law"). Clicking on a post navigates the browser to the corresponding thread.

Settings: Just under the atlas chart, there are several settings you can adjust. The "+" and "-" buttons increase and decrease the size of the focus area, respectively. Depending on the browser you use, you may also be able to adjust the focus area size by scrolling the mouse wheel. Using the "Number of keywords" slider, you can adjust the number of keywords with which the documents within the focus area are described.

Clicking the "ON/OFF" button switches the display of forum posts' titles on and off. By default, titles are hidden and only appear when you move the mouse over a post in order to reduce visual clutter. If you choose to display the titles permanently, their size can be adjusted with the "Font size" slider just next to the "Subjects" button.

3.5. Using the Topical Timeline

To see a timeline of topical evolution of a chosen subpart of the forum as illustrated in Figure 6, first indicate your interest by selecting several threads as described in section 3.2. Then, press the "Timeline" button on the VIDI toolbar.

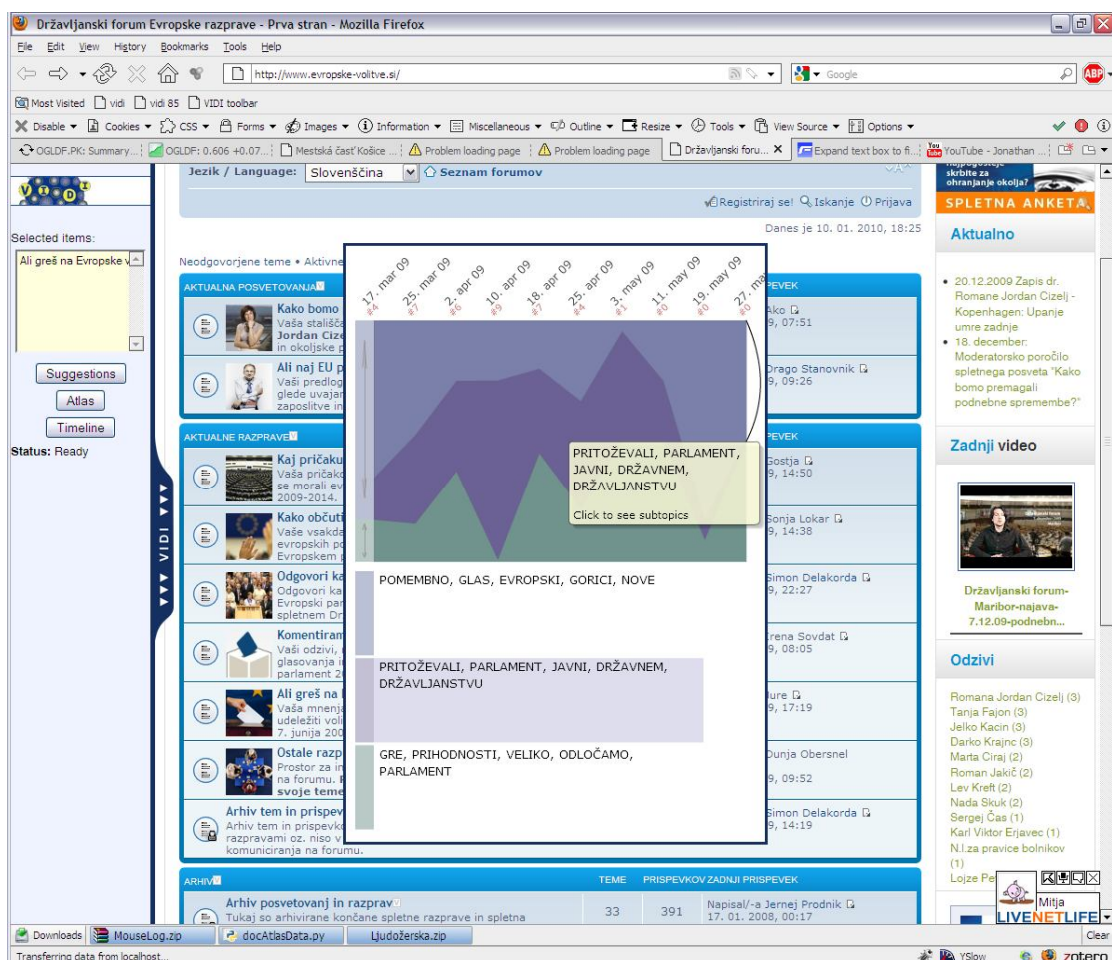


Figure 7: Topical timeline. Each colored stripe represents a topic; its description (in the form of keywords) is given below the graph. The thickness of each stripe corresponds to how much a topic was talked about at a given moment.

The timeline appears in the middle of the window. To discard the timeline, click anywhere outside the popup area.

The main part of the visualization is the graph in its upper half. Each of the colored areas of the graph represents a topic on the forum. The thickness of the colored stripe shows how active this topic was through time. The actual dates are given just above the graph, at the very top of the visualization.

Hovering over a topic shows a tooltip with keywords describing the topic. At the same time, keywords for all the topics are displayed below the graph in a color-coded legend. Additionally, when hovering over a topic, small red numbers are displayed at the top of the graph for each time slot. These are an absolute indicator of the topic's popularity: they represent the actual number of posts from the given time period talking about the given topic.

The topics are automatically determined by semantically clustering the selected posts in a hierarchical fashion. To avoid over-segmentation, all the posts are initially split into just two topics.

If you wish to delve deeper into a topic, click its stripe in the graph. If subtopics are available, the topic will split into two. Click on the black arc on the right to merge the subtopics once again. Click on the vertical arrows in the left margin to temporarily hide all other topics and expand the selected topic (along with its subtopics) over the whole graph. Click on the light grey arrow in the upper left corner to show the remaining topics again.

3.6. Defining Notification Patterns

VIDI-System supports an email based user notification about user-relevant situation within a discussion forum. In order to notify the user he/she must model the situation of interest and register in the VIDI-System. For this reason VIDI notification pattern user interface (UI) provides discussion forum specific categories such that the user can select the relevant categories, configure and connect them to each other using the provided operator nodes. In order to define a notification pattern the user can use the drag and drop functionality provided by the UI.

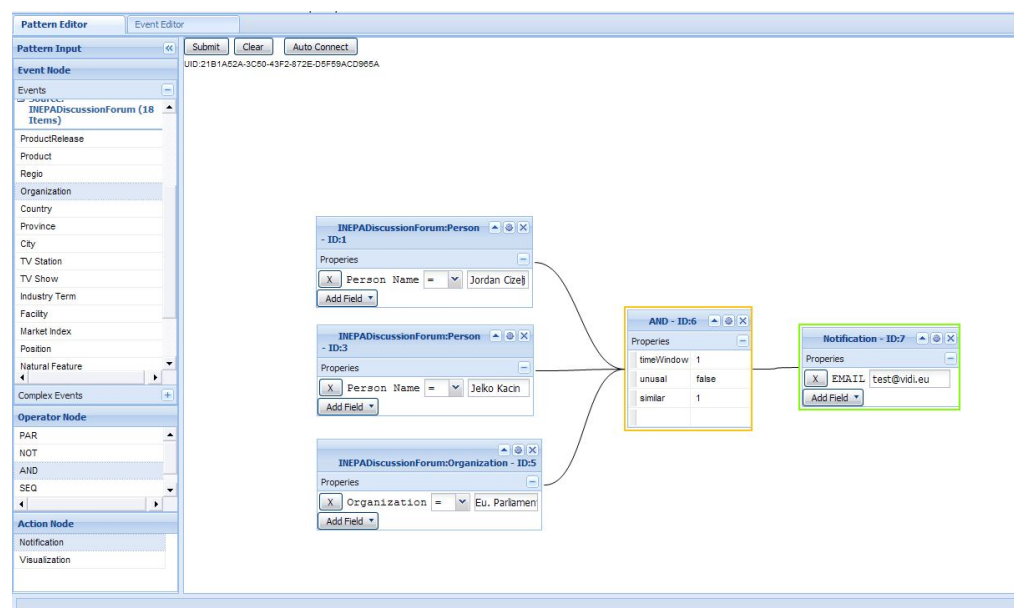


Figure 8: Example of a notification pattern for the INEPA discussion forum

The example in figure 8 shows a pattern that describes a notification pattern within the INEPA discussion forum where two members of European Parliament from Slovenia are mentioned in the European Parliament context.

4. LIVENETLIFE

LiveNetLife is an existing software package which offers automatic connection establishment and real-time chat between users who are browsing completely independent but topically related web pages. As it was considered relevant to boosting e-participation, plans were made in VIDI deliverable D1.1 to adapt it and include it in the VIDI project.

The plans have been followed through and LiveNetLife is now deployed on the Slovene use case web site, www.evropske-volitve.si, and could possibly be added to the remaining use cases as well. Client-side deployment for LiveNetLife will not be offered within the scope of VIDI since LiveNetLife is being developed independently of the project at this stage it does not offer open access to its services to a potentially uncontrollable number of users.

5. TECHNICAL BACKGROUND

The software architecture largely follows the plans proposed in VIDI deliverable D2.1; the main components are therefore only briefly outlined in this section. Consult Appendix A for a list of actual source code files corresponding to architectural components described here.

5.1. Data Acquisition

A local copy of all data from all supported forums is stored in an SQL database. For the Hungarian use case, Blaž Novak has written a specialized web crawler to obtain the data. For the Slovene and Slovak use cases, data is obtained with direct SQL access to the respective databases; only some additional DB schema translation is needed. Both the web crawler and the SQL crawler are run periodically to keep the local copy of the data fresh.

5.2. Data Aggregation and Augmentation

Some established preprocessing steps are performed on all forum posts once they are stored in the database: HTML cleanup, tokenization, lemmatization, stopword removal, frequent n-gram extraction. We adapted an existing lemmatizer (Juršič et al., 2007); Slovene and Hungarian stemming rules were provided by the authors, Slovak ones were trained on the Slovak National Corpus (available at <http://korpus.juls.savba.sk/>). We also track the most frequent surface form for each lemma. Additionally, we perform named entity extraction and consolidation.

After preprocessing and named entity extraction, all distinct terms are enumerated and a sparse vector of term frequencies is stored for each post. To speed up processing of analytic modules, we also store sparse TF vectors for all forum threads and topics (and update them upon post insertions).

We also cache some other basic statistics, e.g. document frequencies for all terms, average post length etc. Some of the caches refresh in real time, some have to be refreshed by periodically running appropriate scripts.

5.3. Analytic Modules

The analytic modules (providing input for browsing suggestions, topical atlas and topical timeline GUIs) access the database directly. The computationally intensive parts are written in C++ and exposed to python with Boost.Python. The python wrapper performs some additional formatting and exposes the functions as web services. Database is accessed with either native drivers (python) or via ODBC (C++).

For browsing suggestions, ranking is performed based on cosine similarity of TF-IDF vectors of the query threads and target threads.

For the topical timeline, hierarchical bisecting k-means with TF-IDF cosine distance is used on posts' sparse TF vectors.

For the topical atlas, the high-dimensional space defined by the sparse TF vectors is projected onto several hundred dimensions using LSI (latent semantic indexing) and from there onto two dimensions using MDS (multidimensional scaling). To speed up the process, the projection is determined by only observing the distances between up to several hundred clusters of documents.

5.4. *Client Side and Client-Server Communication*

The client side of the software is written in Java and snippets of Javascript using the GWT (Google Web Toolkit) platform. The visualizations (topical atlas, topical timeline) are written in Flash with ActionScript 2.

Javascript (the toolbar) and the Flash visualizations communicate using flashvars (from Javascript) and Flash's ExternalInterface class (to Javascript; both directions possible). Flash communicates with the server using custom formatted GET HTTP requests. Javascript communicates with the server using JSONP (JSON with padding) callbacks to work around cross-domain scripting restrictions.

6. REFERENCES

- Fortuna, B., Grobelnik, M. and Mladenić, D.: "Visualization of Text Document Corpus". Informatica Journal 29, 2005, pp. 270-277.
- Grčar, M. 2009, "D2.1: Architecture of the VIDI Integrated System and Test Scenarios ", VIDI Project Report
- Juršič, M., Mozetič, I., Lavrač, N. 2007, "Learning Ripple Down Rules for Efficient Lemmatization". Proceedings of the 10th International Multiconference Information Society, IS 2007, Vol. A, pp. 206-209, Ljubljana.
- Stojanović, N. & Grčar, M. 2009, "D1.1: As-Is Analysis and Tool Selection", VIDI Project Report

APPENDIX A. SOURCE CODE ORGANIZATION

This section is highly technical in nature. It explains the basic folder structure of the various software components comprising VIDI. To make the code package self-contained, this part of documentation is given in a README.TXT file accompanying the source code and is merely repeated here in its original form.

- * /README.txt
This file. Sketches the directory structure.
- * /flash
Flash movies. The movies do not perform any serious computation; data is computed on the server side, the movies merely display it and allow some user interaction.
 - * /flash/docAtlas
Source code for the topical atlas flash movie.
 - * /flash/canyonFlow
Source code for the topical timeline flash movie.
- * /ForumExplorerBar
The GWT project for the client side of the VIDI toolbar.
 - * /ForumExplorerBar/src
Java sources (which then get compiled to javascript by GWT). Also, some .xml configuration files for the project.
 - * /ForumExplorerBar/war
Additional resources: css for the toolbar and some more .xml config files. Most resources are in /server/gfx, though. Note the .js file -- this is a hacked version of what GWT produces, with instructions on how to reapply the hack to future GWT outputs. The hack enables to use the toolbar as a bookmarklet, something which is not possible by default.
- * /server
Server side of the toolbar. Implemented in python, with the heavy computation offloaded to C++ libraries.
 - * /server/glib
JSI's C++ standard template libraries and text mining libraries.
 - * /server/clib
A small C++ project which produces a dll to be used by python. Uses glib for algorithmics and Boost.Python for exposing functions and classes.
 - * /server/gfx
Images referenced by the toolbar.
 - * /server/json
A python library for handling JSON data.
 - * /server/{canyonFlow,docAtlas}Data.py
Scripts for generating XML inputs for the flash movies. These act as simple web services (with their own parameter syntax)
 - * /server/crossdomain.xml
This file must be here to allow flash movies from accessing the .py web services from the previous bullet point.

- * /server/db_triggers.py
Database triggers for maintaining up-to-date word statistics. The triggers are written in python for postgres.
- * /server/index.py
A web service mini-framework (+ accompanying services) for handling arbitrary function calls from the toolbar. In the end, there are only two services present, notably one for computing browsing suggestions.
- * /server/structure.py
A library for "structuring" the database -- computing and extracting named entities, updating statistics etc. Also contains many standalone functions for outstanding maintenance work on the database (e.g. complete recomputation of some statistic).
- * /server/sync*
Scripts for obtaining data from forums, either via SQL connections or by crawling, and caching it in the local database. Uses structure.py.